DOCUMENT RESUME

ED 194 582 TM 800 711

AUTHOR Wing, Hilda
TITLE Age, Sex, and Repetition Effects with an Abilities Test Battery.
INSTITUTION Office of Personnel Management, Washington, D.C.
PUB DATE Sep 80
NOTE 19p.: Paper presented at the Annual Meeting of the American Psychological Association (88th, Montreal, Quebec, Canada, September 1-5, 1980).
JOURNAL CIT Applied Psychological Measurement; v4 n2 p141-55 Spr 1980

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Age Differences; *Cognitive Tests; College Graduates; Federal Government; Higher Education; Personnel Selection; *Scores; *Sex Differences; Test Interpretation; *Test Wiseness
IDENTIFIERS *Professional and Administrative Career Exam; *Repetition Effects

ABSTRACT

Prior research with the Professional and Career Administrative Examination (PACE), a multiple abilities test battery for federal candidates, demonstrated practice effects (score gains) on alternate forms of the test parts within one administration of the battery. These effects were largest for those abilities including item types constructed according to explicit rules (induction, deduction), next largest for abilities subject to speededness (number), and least for those abilities tapping general information (verbal, judgment). The present study again demonstrated practice effects by ability--over several administrations. About 48,000 candidates, mostly college seniors and graduates, were identified as having taken the PACE two to five times between 1974 and 1977. The most important contribution to score gains was the number of prior testings. The largest gains were from the first to the second testing. The total test score gain for repetition (0.3 standard deviation) was greater than that predicted from practice effects alone. The repetition gain for number was less than the practice gain, indicating the ephemeral nature of score gains where speededness is involved. Higher initial scores were earned by those repeating the battery the fewest times; higher final scores were earned by those repeating the most times. Age and sex were related to initial score level but not to score gain; age differences in test scores were greatest for those test parts showing the largest effects of practice or repetition. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194582

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Age, Sex, and Repetition Effects with an Abilities Test Battery

Hilda Wing

Office of Personnel Management

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Wing

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the 88th convention of
the American Psychological Association
Montreal, Quebec, Canada

September, 1980

TM 800711

The opinions expressed herein are those of the author and do not necessarily reflect those of the Office of Personnel Management.

Age, Sex, and Repetition Effects with an Abilities Test Battery

Abstract

A multiple abilities test battery (Verbal, Judgment, Induction, Deduction, Number) has been used since 1974 in a nationwide examination for selection into Federal professional and administrative occupations. Prior research demonstrated the existence of practice effects from one alternate form of an ability test part to the next within one administration of the test battery. These effects were largest for those test parts including item types constructed according to explicit rules (Induction, Deduction), next largest for test parts subject to speededness (Number), and least for those test parts tapping general information (Verbal, Judgment). During the first three years of the examination alternate forms of the test battery were administered 17 times to over 600,000 individuals. Somewhat less than ten percent of these people repeated the test battery two or more times. The major interest of this study was whether the practice effects would replicate for test repeaters. Also investigated were possible moderating effects of the number of repetitions as well as the age and sex of the test taker. Test score gains from repetition followed the same pattern as test score gains from practice: Induction and Deduction gains were greater than Verbal and Judgment gains. The total test score gain for repetition (.3 standard deviation) was greater than that predicted from the practice effects research, indicating more about the self-selection factors operating than about any observed variable. However, the repetition gain for Number was less than the practice gain, indicating the ephemeral nature of score gains when speededness is involved. Higher initial scores were earned by those repeating the battery the fewest times; higher final scores were earned by those who repeated the battery the most times. While age and sex did appear to influence the initial level of test scores, these variables did not have any sizeable impact on test score gain. The age differences in test scores were greatest for those test parts showing the largest effects of practice or repetition.

Age, Sex, and Repetition Effects with an Abilities Test Battery

Standardized testing has been under fire, one reason being the public's uncertainty of the stability of test scores. A number of studies have demonstrated that coaching does not improve scores on the Scholastic Aptitude Test (SAT; Donlon & Angoff, 1971; Fremer & Chandler, 1971). However, the Federal Trade Commission (Notes 1 and 2) reported that coaching schools improve SAT scores of individuals whose initial test scores were below those expected based on high school academic performance.

Test score stability became a concern to this author with the equating of several alternate forms of a multiple cognitive abilities test battery used in examining for entry-level professional occupations in the Federal service (Wing, 1980a). Two alternate forms of one of the five tests of a battery (Verbal, Judgment, Induction, Deduction, Number) were administered to a random sample of applicants. Small but consistent score increases from the first test form to the second were observed. Subsequent analyses confirmed that these increases were due to practice. They were greatest for item types constructed according to specific rules, such as in Induction and Deduction, next largest for test parts subject to speededness, such as Number, and least for item types tapping general information, such as in Verbal and Judgment.

A number of research studies on practice effects have been published during this century, the most comprehensive review being nearly forty years

old (Greene, 1941). Thse studies provide corroboration of practice efffects in general and of the influence of specific item types in particular. The public, especially applicants for Federal employment via this examination, could benefit from such information. However, data from randomly selected samples at one test administration having practice on one part of a multiple abilities battery might not be the same as data from self-selected repeaters at two or more test administrations on all parts of the battery. Further, the practice effects data files included no demographic information but a repeater file could include variables such as sex, age, college major, amount of education, and date(s) of testing.

A second reason for contemporary attacks on standardized testing is that different demographic groups have different average scores (Wing, 1980b). The existence of practice effects, however, suggests that such differences may require reinterpretation. Racial/ethnic membership was not routinely collected from applicants for Federal employment until 1978. Sex and age information were available so an empirical base could be obtained concerning the relationships of these independent variables to test scores.

Between November, 1974, and May, 1977, six alternate forms of each part of a multiple abilities test battery were used, in various combinations, at 17 nationwide test administrations. Over 600,000 individual testings occurred; 48,396 individuals were identified as having taken the test battery twice; nearly all (over 99%) had taken it five or fewer times. This report discusses the changes in test scores as a function of cognitive ability tested, sex, and age as indexed by year of birth, separately for those repeating two, three, four, or five times.

Method

Research Participants

Of the over 600,000 applicants between November, 1974, and May, 1977, approximately 60% were male, over 80% were college seniors or graduates, and less than 30% were over 30 years old. Repeaters were identified by matching test records on name and date of birth. Over 97% of the 48,053 identified as taking the battery from two to five times had provided valid sex information. Age was coded from year of birth into seven categories; 41 individuals assigned to the oldest category (born before 1900) were dropped from further analyses. The data of this report were provided by 24,528 males and 14,469 females taking the test twice ("Twos"), 3,979 males and 1,900 females taking the test three times ("Threes"), 983 males and 420 females taking the test four times ("Fours"), and 347 males and 96 females taking the test five times ("Fives").

Abilities Assessed

The test battery assessed five abilities as described by French (1951) and by French, Ekstrom, and Price (1963).

Verbal: vocabulary and reading items.

Judgment: comprehension items, in each of which a paragraph was presented and the most plausible of the five alternatives was to be selected. The solution tapped general knowledge not included in the original paragraph.

Induction: letter series and geometric analogies.

Deduction: tabular completion (missing values in tables or charts had to be deduced from the remaining information) and inference items (a statement

or paragraph was to be accepted as true and the correct answer was to be derived from the statement without drawing on outside information).

Number: arithmetic computation and arithmetic reasoning items, written at about the eight grade level and designed to minimize the importance of verbal ability.

Practice effects had appeared to be greatest for Induction, then Deduction and Number, and minimal for Verbal and Judgment. Analyses of omit rates indicated that part of the effect for Number could be attributed to pacing: when omits were excluded from test scores the practice effect for Number was reduced (Wing, 1980a).

Test Materials and Procedure

The written test battery consisted of six separately timed parts each requiring 35 minutes. Five of these, one for each ability, counted towards an individual's employment rating and consisted of 30 items each. Each of the five raw test scores was converted to the same scale (Notes 3 and 4).

Data Analyses

Unweighted means analyses of variance of Sex by Age by Test Part, the last a repeated factor, were used to evaluate test scores (Winer, 1971). These analyses were performed using the number of test repetitions as an additional between subjects variable, based on the initial and final test scores. Additional analyses compared the first two sets of test scores for all four groups of repeaters, the first three for three groups, the first four for the Fours and Fives. Such analyses evaluated the similarity of the four groups of repeaters.

Other ANOVAs were performed separately for each repeater group, on the varying number of test score sets as dependent variables. To the extent that all these analyses replicate each other, they corroborate the analyses of each repeater group.

Results and Discussion

Analyses of variance across all repeater groups support the following conclusions.

1. The most important contribution to test score variance, for each repeater group, is the number of prior testings. The largest gains were from the first testing to the second.

2. The next most important factor, for each repeater group, was type of ability. The largest gains were for Induction, then Deduction, then Number, then Verbal and Judgment.

3. The interaction of Test Part and Number of Testings was also significant for all groups. Although the increases for each test part diminished after the first, this decrease was greater, the larger the initial practice effect. That is, the larger gains diminish faster than the smaller.

4. Females had different average scores than males. They had higher averages in Induction, the same in Verbal and Deduction, and lower in Judgment and Number.

5. Age was related to test scores for three of the groups. Age was not significant for the Fives, but the interaction of Age by Sex by Number of Testings was. Age by Sex was significant for all but the Fours. The older repeaters had lower averages than the younger; the younger males had higher

averages than the younger females while the older females, with the exception of the oldest Fives, had the same or higher averages than the males.

6. Age was related to test part in that the difference between younger and older was most pronounced for Induction and Deduction, slight for Number, and negligible for Verbal and Judgment.

7. Age did not interact with number of testings; Sex interacted with Number of Testings only for the Twos. That is, score gains depend on the ability and the number of prior testings but not on the age or sex of the test-taker.

8. A final conclusion appears robust: The number of repetitions was related to both initial and final scores. The fewer the repetitions, the higher the initial score; the more the repetitions, the higher the final score. This can be seen in Table 1, which displays the mean scores for each testing by age and sex for each repeater group.

Inferences must be qualified by stressing possible factors which led individuals to select themselves to take a 3-1/2 hour test battery two to five times over a three year period. These individuals might have believed their initial scores underrepresented their ability. Many of them may have participated in exercises specifically designed to improve scores in this battery. And, of course, their initial scores were probably insufficiently high for employment selection.

The pattern of repeater effects is the same as previously observed for practice effects (Wing, 1980a), with two exceptions: the actual increases appear larger for repeaters while the relative increase for Number appears

smaller. There is no ready explanation for the first exception beyond repeating the limitations to generalizations, stated above. For the second, recall that a change in pacing was hypothesized to have contributed part of the practice effect for Number. It is plausible that adjusting one's rate of work for a specific test part is more labile than adjusting one's style of work to solve given types of test questions.

One interesting finding of this study is the similarity between the relation of type of ability to score gain and the relation of type of ability and age to initial score: Age differences were greatest for those tests showing the greatest practice effects. However, the change from initial score appeared to be unrelated to either age or sex. It would be important to replicate these findings in other contexts.

Reference Notes

1. Federal Trade Commission, The effects of coaching on standardized admission examinations. Boston: Author, 1978.

2. Federal Trade Commission. Revised statistical analyses of data gathered by Boston regional office of the Federal Trade Commission. Washington, D.C.: Author, 1979.

3. Wing, H. Scaling and transmutation procedures, Test 500. Washington, D.C.: U.S. Civil Service Commission, 1975. Operations Paper.

4. Wing, H. Scaling and equating series of Test 500, FY 75. Washington, D.C.: U.S. Civil Service Commission, 1975. Operations Paper.

References

Donlon, T. F. & Angoff, W. H. The Scholastic Aptitude Test. In Angoff, W. H. (Ed.) The College Board Admissions Testing Program. New York: College Entrance Examination Board, 1971, 15-47.

Fremer, J. & Chandler, M. O. Special studies. In Angoff, W. H. (Ed.) The College Board Admissions Testing Program. New York: College Entrance Examination Board, 1971, 147-181.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. Psychometric Monographs, 1951, No. 5.

French, J. W., Ekstrom, R. B., & Price, L. A. Manual for kit of reference tests for cognitive factors. Princeton, N.J.: Educational Testing Service, 1963.

Greene, E. B. Measurement of human behavior. New York: Odyssey Press, 1941.

Winer, B. J. Statistical principles in experimental design (Rev. ed.) New York: McGraw-Hill, 1971

Wing, H. Practice effects with traditional mental test items. Applied Psychological Measurement, 1980, ~~in press. (a)~~ 4, 141-155.

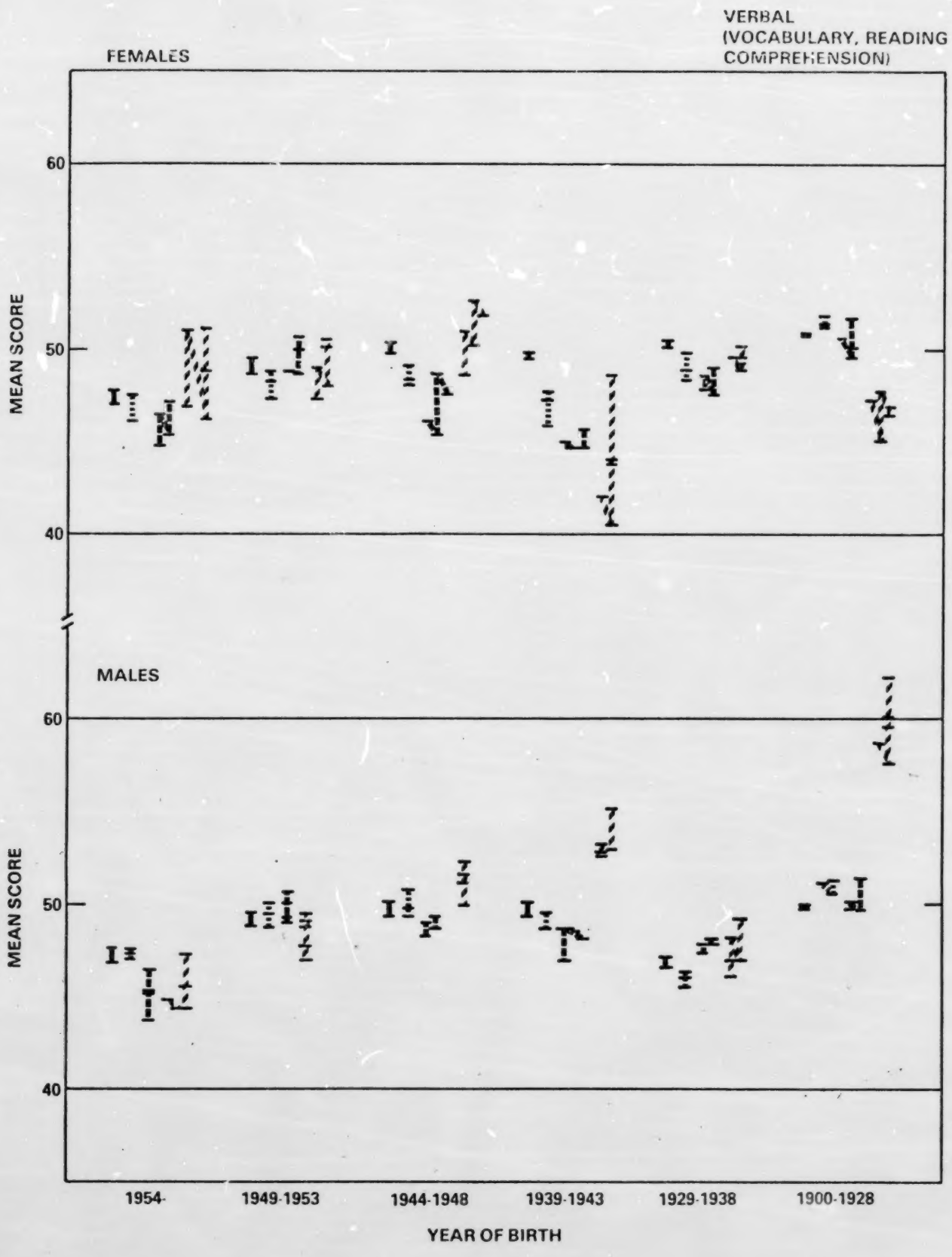
Wing, H. Profiles of cognitive ability of different racial/ethnic and sex groups on a multiple abilities test battery. Journal of Applied Psychology, 1980, 65, 289-298. (b)

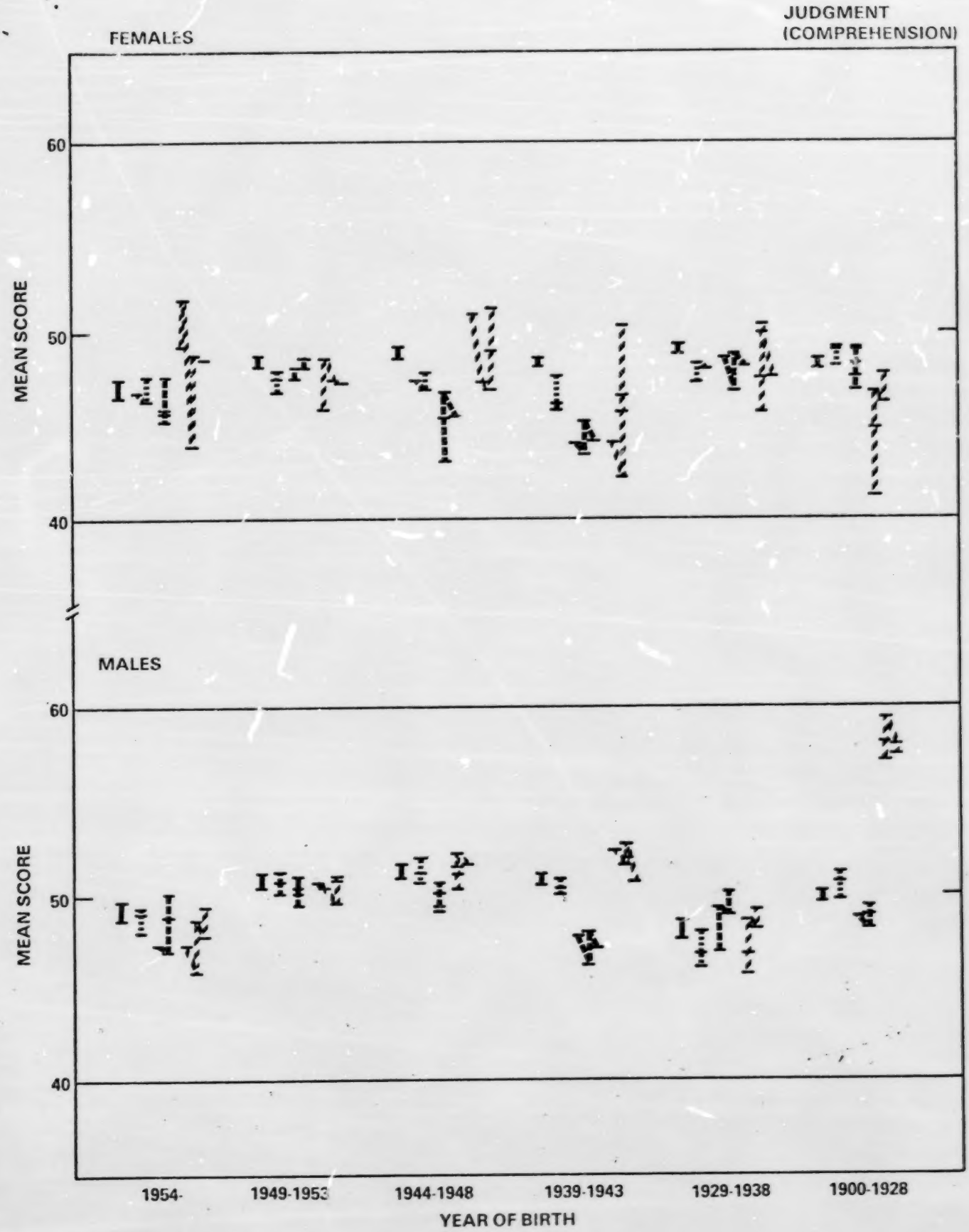
Table 1

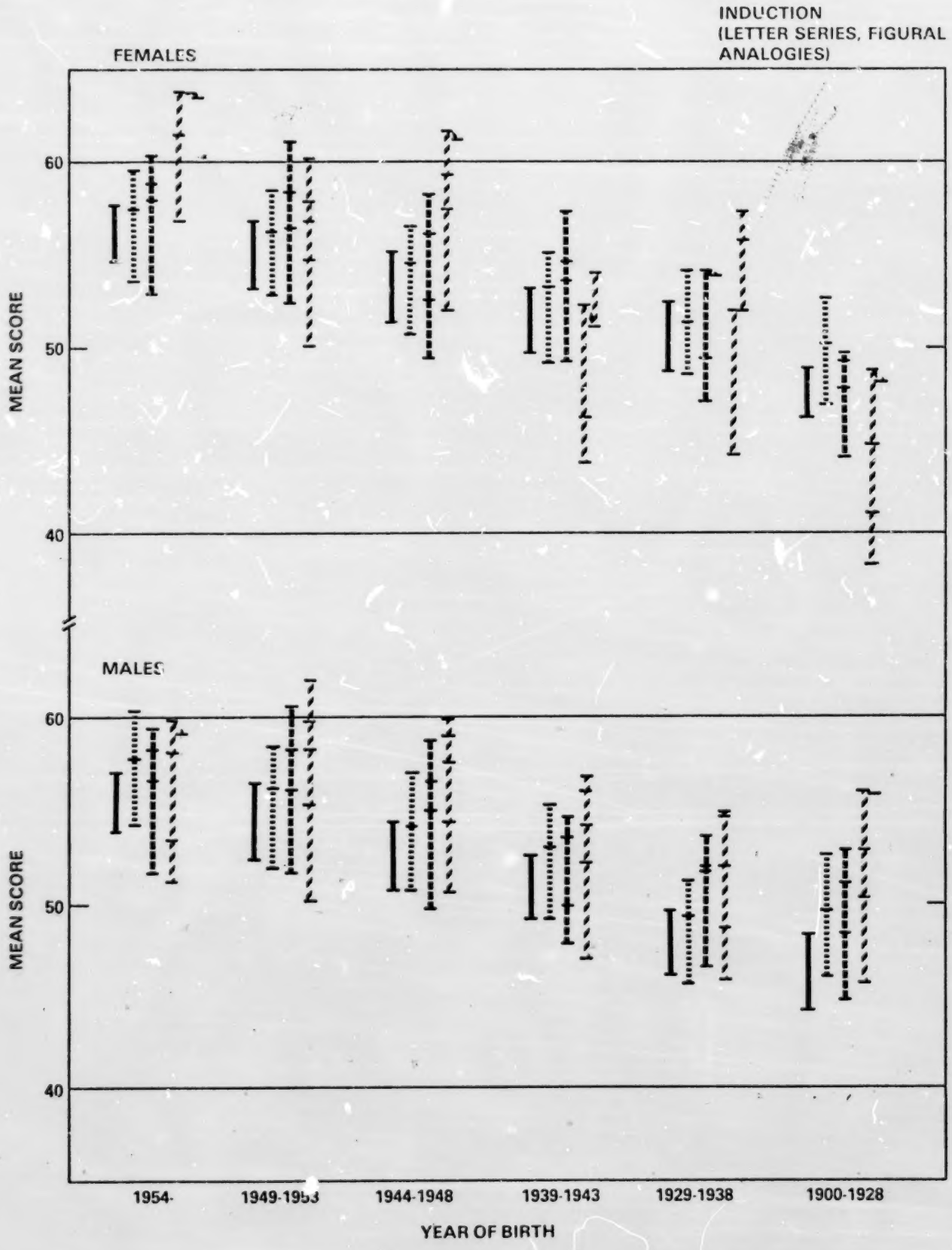
Mean Test Score for Each Testing by Age and Sex for Four Repeater Groups

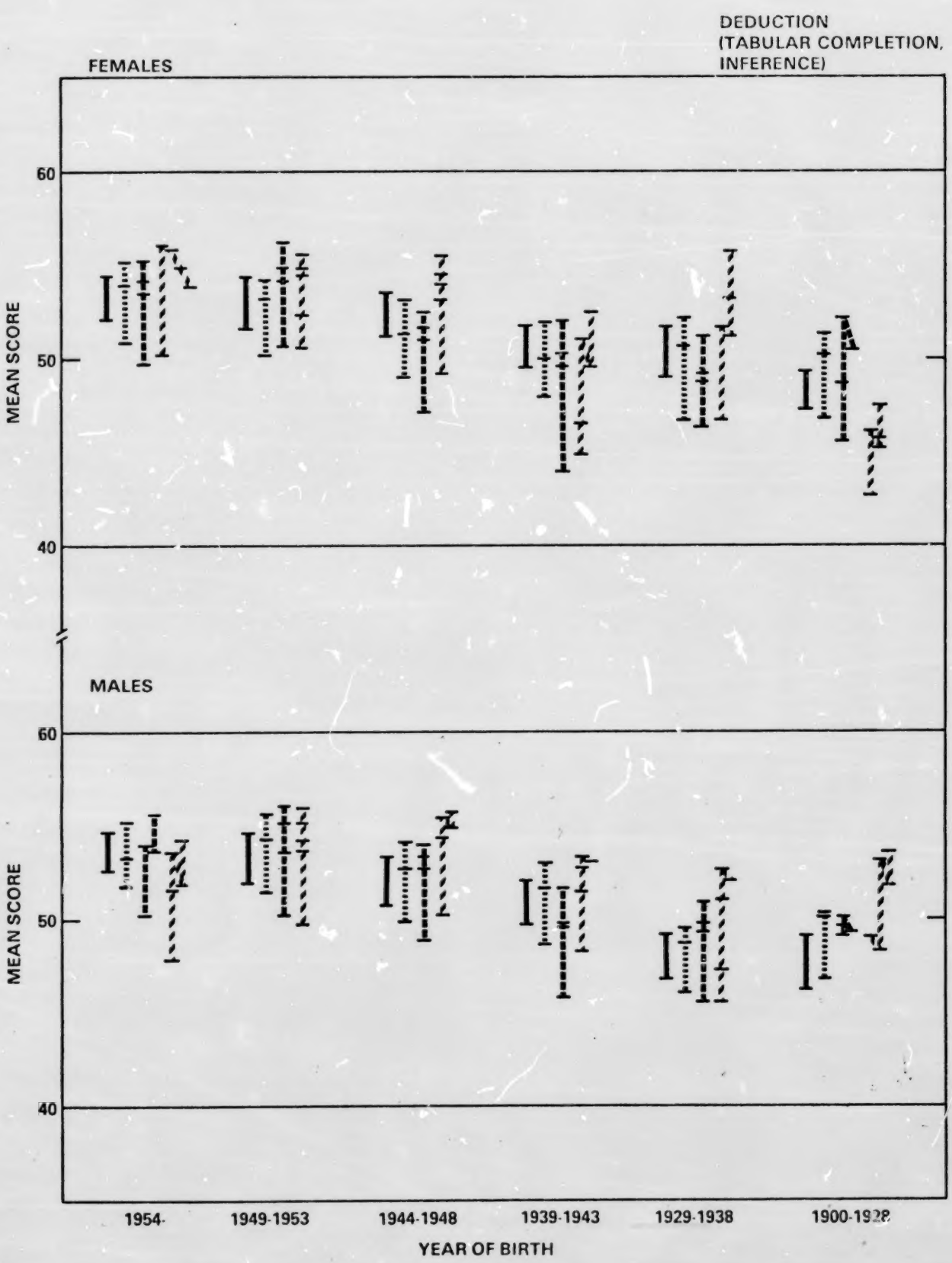
	Sex										
	Males					Females					
	N	Test number					N	Test Number			
		1	2	3	4	5		1	2	3	4	5
1954-1970												
Twos	2,134	50.8	52.6				2,171	49.9	51.7			
Threes	274	50.4	52.0	53.3			220	49.1	51.0	52.0		
Fours	49	48.9	51.1	51.8	52.9		44	48.4	50.7	50.9	52.4	
Fives	17	48.0	49.1	51.6	51.6	53.0	4	51.2	54.5	52.7	54.6	53.9
1949-1953												
Twos	12,026	51.1	53.2				7,264	50.2	52.2			
Threes	1,967	50.8	52.9	54.1			941	49.1	51.1	52.1		
Fours	497	50.3	52.4	53.7	54.7		204	49.7	51.8	52.8	53.8	
Fives	161	49.7	51.9	52.8	53.8	54.5	43	48.2	50.3	51.6	52.4	53.0
1944-1948												
Twos	5,989	50.7	52.7				2,127	50.1	52.0			
Threes	994	50.4	52.3	53.6			244	48.7	50.4	51.5		
Fours	255	49.4	51.7	52.4	53.4		64	46.7	48.8	50.9	50.8	
Fives	103	50.6	52.8	54.1	54.7	55.1	19	49.8	51.6	52.3	54.7	54.8
1939-1943												
Twos	1,552	50.0	51.7				849	48.9	50.7			
Threes	278	49.4	51.4	52.3			145	47.4	49.4	50.7		
Fours	74	47.5	48.8	50.3	50.4		35	45.7	48.1	48.6	50.1	
Fives	36	49.5	52.1	52.8	53.5	53.4	8	44.4	44.5	48.6	47.4	52.0
1929-1938												
Twos	1,893	47.3	49.0				1,190	48.9	50.8			
Threes	329	46.4	48.3	49.1			183	47.6	49.5	50.9		
Fours	62	47.0	49.8	49.8	51.2		36	47.4	48.6	49.6	50.1	
Fives	21	46.6	48.3	50.9	51.2	52.0	11	47.1	50.5	51.0	52.0	53.0
1900-1928												
Twos	934	47.9	49.7				868	47.9	49.4			
Threes	137	48.7	50.5	51.5			167	47.9	50.0	51.1		
Fours	46	48.4	49.6	50.4	51.2		37	46.8	48.5	50.1	50.6	
Fives	9	52.9	53.4	55.4	55.6	56.7	11	42.4	43.9	45.6	46.5	47.4

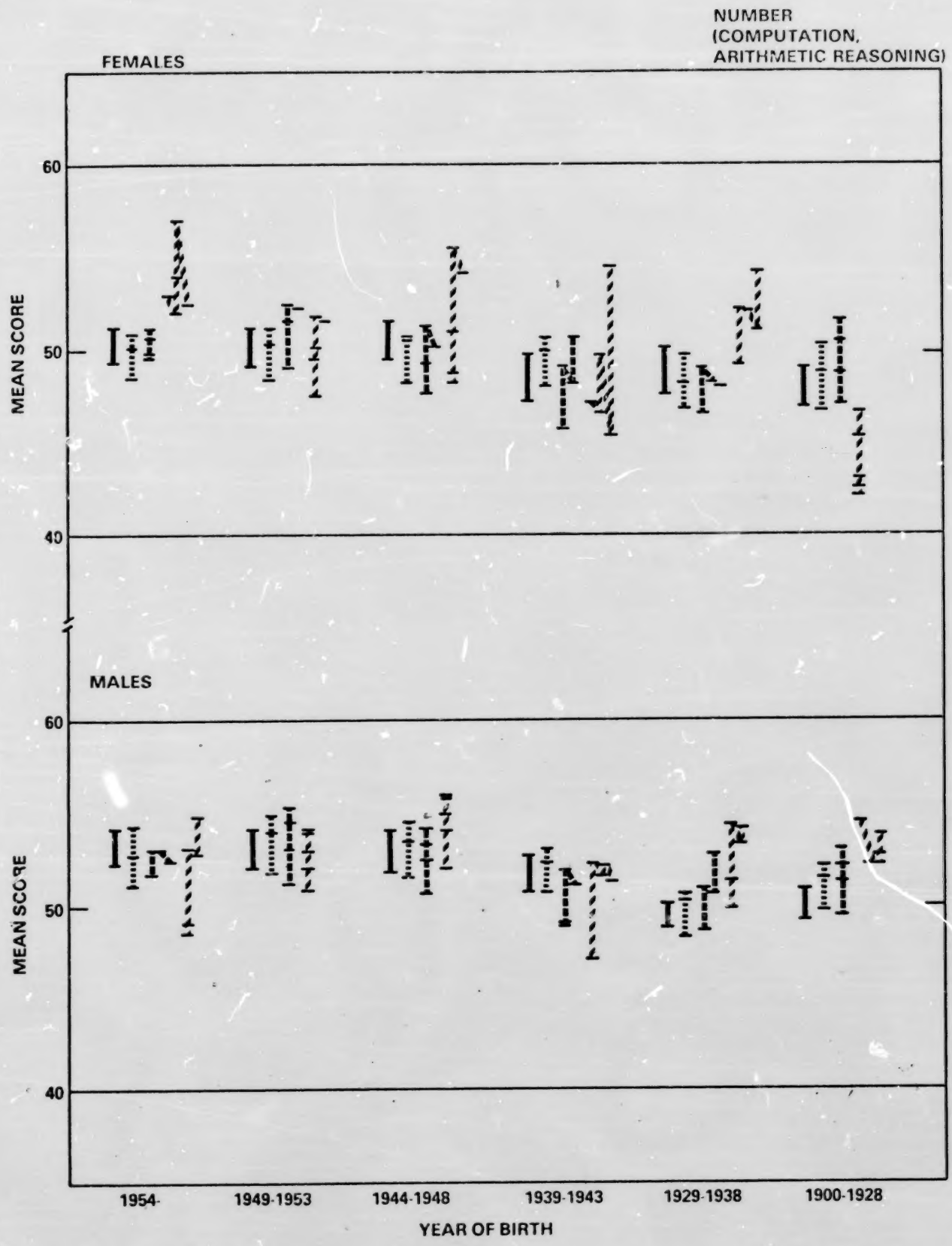
Note: Test part scores equated to scale with mean of 50 and standard deviation of 10. The standard deviation of these averaged totals scores is less than 8.

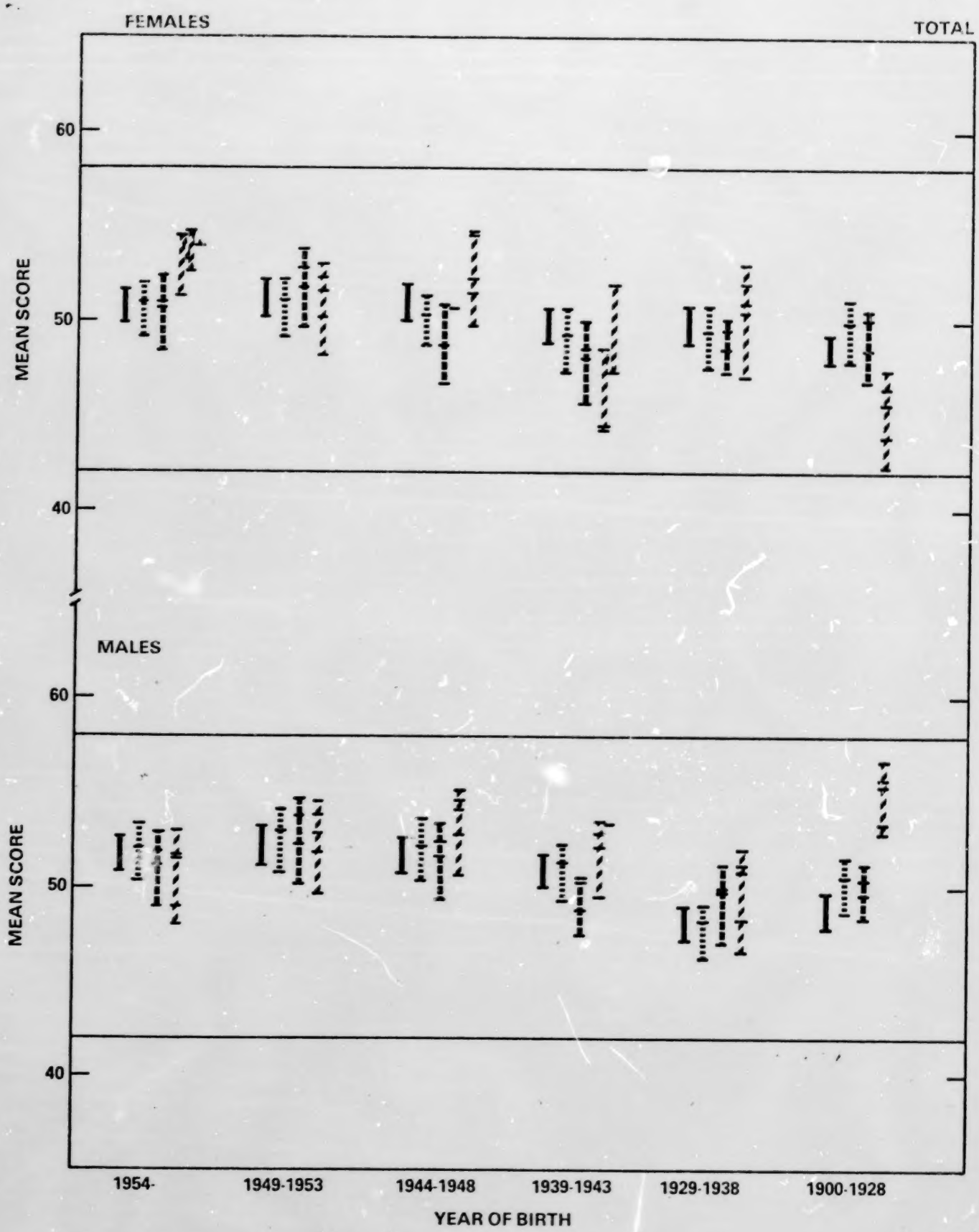

VERBAL
(VOCABULARY, READING
COMPREHENSION)
FEMALES
MEAN SCORE
60
50
40
MALES
MEAN SCORE
60
50
40
1954-
1949-1953
1944-1948
1939-1943
1929-1938
1900-1928
YEAR OF BIRTH


JUDGMENT
(COMPREHENSION)
FEMALES
MEAN SCORE
60
50
40
MALES
MEAN SCORE
60
50
40
1954-
1949-1953
1944-1948
1939-1943
1929-1938
1900-1928
YEAR OF BIRTH


INDUCTION
(LETTER SERIES, FIGURAL ANALOGIES)
FEMALES
MALES
MEAN SCORE
MEAN SCORE
60
50
40
60
50
40
1954-
1949-1953
1944-1948
1939-1943
1929-1938
1900-1928
YEAR OF BIRTH


DEDUCTION
(TABULAR COMPLETION,
INFERENCE)
FEMALES
MALES
MEAN SCORE
60
50
40
1954-
1949-1953
1944-1948
1939-1943
1929-1938
1900-1928
YEAR OF BIRTH


NUMBER
(COMPUTATION,
ARITHMETIC REASONING)
FEMALES
60
50
40
MEAN SCORE
MALES
60
50
40
MEAN SCORE
1954-
1949-1953
1944-1948
1939-1943
1929-1938
1900-1928
YEAR OF BIRTH


FEMALES
TOTAL
60
50
40
MEAN SCORE
MALES
60
50
40
MEAN SCORE
1954-
1949-1953
1944-1948
1939-1943
1929-1938
1900-1928
YEAR OF BIRTH